TWO SYNTHESIS METHODS BASED ON CEPSTRAL PARAMETERIZATION

Jiří PŘIBIL[1], Anna MADLOVÁ[2]

[1] IREE CAS, Chaberská 57
182 51 Praha 8, Czech Republic

[2] KRE FEI STU, Ilkovičova 3
812 19 Bratislava, Slovak Republic

Abstract

The paper deals with two implementations of the speech synthesis based on the cepstral representation of the human vocal tract model. Both the approaches to speech modeling are evaluated in the frequency domain. The paper also compares computational complexity of these two methods.

Keywords

Signal processing, speech processing, speech analysis and synthesis

1. Introduction

Linear predictive coding (LPC) is often used for speech modeling. The real cepstrum coding (RCC) is parametrical and comprises not only formants but also antiformants of the speech spectrum in contrast to the LPC model, which only realizes formants by the all-pole transfer function [1]. It does not apply any simplifying assumptions about the speech production system and contains also information about the spectrum of the vocal tract excitation.

Harmonic speech modeling represents the speech signal as a sum of harmonically related sine waves with frequencies given by pitch harmonics, and amplitudes and phases given by sampling the vocal tract model transfer function at these frequencies. The vocal tract may be represented by LPC or real cepstrum. In this paper cepstral parameterization is used for comparison of source-filter and harmonic speech modeling.

2. Cepstral speech analysis

The cepstral speech model corresponds to the resynthesis of the speech spectrum comprising poles as well as zeros of the model transfer function and containing also information about the spectrum of the model excitation. The cepstral speech synthesis is performed by a digital filter implementing approximate inverse cepstral transformation. The system transfer function of this filter is given by an exponential relation

$$G(z) = \exp\left[\tilde{S}(z)\right] \tag{1}$$

where exponent is the Z-transform of the truncated speech cepstrum

$$\tilde{S}(z) = \sum_{n=0}^{N_0} \tilde{s}_n z^{-n}, \tag{2}$$

and $\{\tilde{s}_n\}$ represents the minimum phase approximation of the real cepstrum $\{c_n\}$

$$\begin{aligned} \tilde{s}_n &= c_n, && n = 0, N_{FFT}/2, \\ \tilde{s}_n &= 2c_n, && 1 \le n < N_{FFT}/2, \\ \tilde{s}_n &= 0\ , && N_{FFT}/2 < n \le N_{FFT} - 1, \end{aligned} \tag{3}$$

where the N_{FFT} is the number of points of FFT. The system transfer function is defined as

$$G_F(z) = \exp(c_0) \prod_{i=1}^{N_0} G_i(z), \tag{4}$$

and can be performed by a cascade connection of N_0 elementary filter structures.

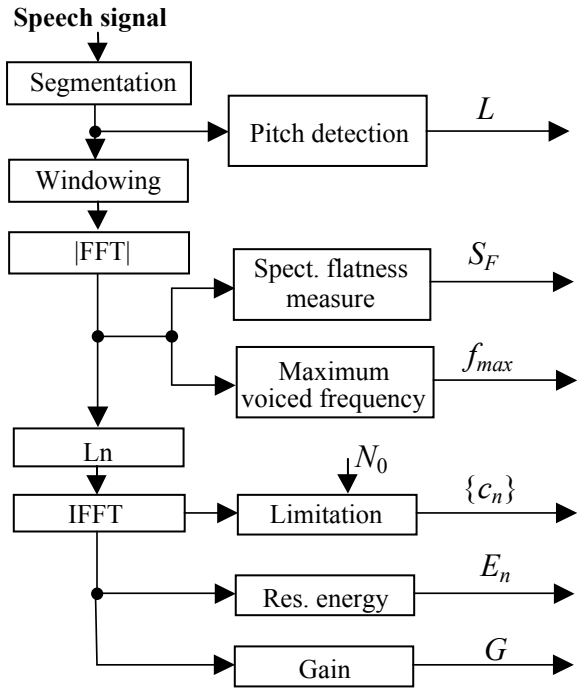

Fig. 1 Block diagram of the cepstral analysis.

The error of this approach is caused by two sources: the limited number of the used cepstral coefficients and the applied approximation of the inverse cepstral transformation, which can be performed by Padé approximation of the exponential function. It has been found out by simulation that the minimum number of N_0 (25 at an 8-kHz rate, 50 at a 16-kHz rate) cepstral coefficients is necessary for sufficient log spectrum approximation [2]. The residual signal energy E_{res} from the limited real cepstrum is defined as

$$E(z) = \exp\left[2 \sum_{n=N_0+1}^{N_{FFT}/2} c_n z^{-n} \right]. \quad (5)$$

The residual signal energy E_{res} is obtained from real cepstrum of the residual signal by FFT and exponentialization

$$E_{res} = \sqrt{\frac{2}{N_{FFT}} \sum_{k=1}^{N_{FFT}/2} \exp\left(E_k^2\right)}, \quad (6)$$

where $\{E_k\}$ is the real part of the spectrum of the limited real cepstrum [3]. The result energy *En* for cepstral synthesis is given as a multiplication of $\exp(c_0)$ * E_{res} (Fig. 4).

3. Synthesis pitch synchronization

For cepstral speech analysis we have used the signal processing approach, which is based on pitch asynchronous segmentation with the length N=192/384 samples (f_s=8/16 kHz) which are corresponding to the 24 ms frames. However, the speech resynthesis must be performed pitch synchronously to obtain the right length proportionality between voiced and unvoiced parts in the original and the resynthesized signal and to minimize transient effect.

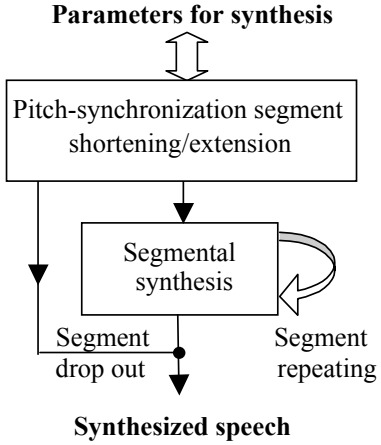

Fig. 2 Synthesis pitch synchronization method.

In the pitch synchronous synthesis the length of the segment *NSYNT* is controlled by a parameter *NVYROV* the value of which is defined as the difference *NSYNT-N/2* [4]. In the parameter *NVYROV* the differences from the synthesis of the previous segments are accumulated. In the case of a voiced segment the length is defined as the fix-point multiple of the pitch- period L: $NSYNT = L*i$, i = 0, 1, … For unvoiced segment the length is given by the difference *N/2-NVYROV* (segment shortening or extension). The standard synthesis with the length *NSYNT* is performed while $|NVYROV| \le N/4$, else the synthesis depends on the polarity *NVYROV*. When $NVYROV < 0$, the segment is dropped out and the new value is $NVYROV = NVYROV + NSYNT$. When $NVYROV > 0$, synthesis of the actual segment is repeated and $NVYROV = NVYROV - NSYNT$.

4. Cepstral speech synthesis

The cepstral synthesis block structure is given by a cascade of digital filters each of which performs the inverse transformation of one cepstral component. The error of this inverse cepstral approximation depends on the number and the values of the applied cepstral coefficients and the approximation structure used [2]. For realization, the exponential function $\exp(\tilde{s}_n z^{-n})$ can be replaced by a rational approximation and we obtain the elementary filter transfer function in the form

$$G_1(z) = \frac{2+x}{2-x}, \quad G_2(z) = \frac{12+6x+x^2}{12-6x+x^2}, \quad (7)$$

$$G_3(z) = \frac{120+60x+12x^2+x^3}{120-60x+12x^2-x^3}, \quad x = \tilde{s}_n z^{-n}.$$

The transfer functions of these elementary filters are given by the 1st, 2nd or 3rd order Padé approximations implemented in the second canonic form, as shown in Fig. 3.

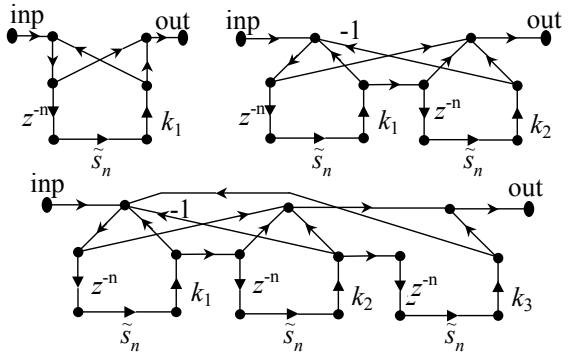

Fig. 3 Computing diagrams of elementary approximation filter structures.

The elementary approximation filter stability criterion and corresponding filter coefficients k_1, k_2, k_3 are shown in Tab 1.

Phonetic research of Czech and Slovak sounds shown that some vowels and voiced consonants contain (besides the voiced excitation) also a high frequency noise component. Therefore, experiments have been performed to determine the voiced/unvoiced energy ratio from the spectral flatness measure S_F which can be estimated during the cepstral speech analysis

$$S_F = \frac{\exp(2c_0)}{r_0}, \quad 0 \le S_F \le 1, \quad (8)$$

where the value $\exp(2c_0)$ represents the square of the geometric mean of the spectrum and r_0 (the zero autocorrela-

tion coefficient) is the energy of the speech segment [2]. According to the statistic analysis of the Czech and Slovak words the ranges of $S_F = (0 \div 0.12)$ for voiced sounds and $S_F = (0 \div 0.65)$ for unvoiced sounds have been estimated.

Filter type	Stability criterion	Filter coefficients		
		k_1	k_2	k_3
1st order	$\lvert\tilde{s}_n\rvert \le 0.75$	1/2	- - -	- - -
2nd order	$\lvert\tilde{s}_n\rvert \le 1$	1/2	1/6	- - -
3rd order	$\lvert\tilde{s}_n\rvert \le 2$	1/2	2/10	1/12

Tab. 1 Filter condition for the 1st, 2nd and 3rd order approximation structures.

To apply this approach to the cepstral speech synthesis, a modification of the excitation signal preparing phase must be executed. The change is practised for excitation signal generation only for the synthesis of voiced sounds. The high frequency noise component is added to the periodical signal from the impulse generator. The mutual proportion between both generated signals is determined by two parameters K_U and K_V defined for voiced segments

$$K_U = \frac{\sqrt{S_F}}{\sqrt{S_F} + \sqrt{1 - S_F}}, \qquad K_V = 1 - K_U \tag{9}$$

and for unvoiced segments

$$K_U = 1 \qquad K_V = 0.$$

The high frequency noise component is produced by the basic random noise generator, whose signal output is high-pass filtered

$$G_{HP}(z) = k_{HP} \cdot \left(1 - z^{-1}\right), \quad k_{HP} = \tfrac{1}{2}. \tag{10}$$

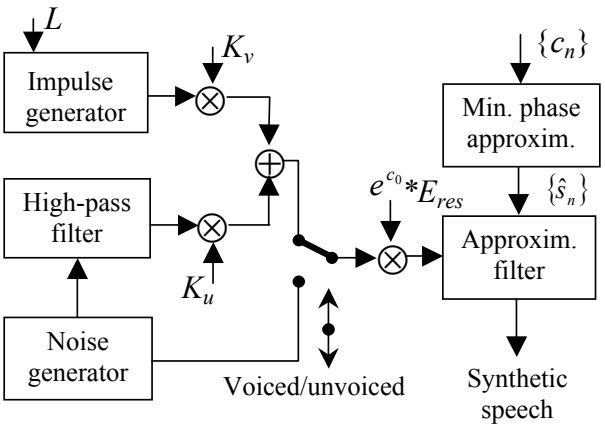

Fig. 4 Cepstral speech synthesizer with the mixed excitation.

5. Harmonic speech synthesis

The harmonic synthesizer [5] with cepstral description [6] utilizes the same number N_0 of cepstral coefficients $\{c_n\}$ as the cepstral synthesizer. These coefficients are used for computation of amplitudes $\{A_m\}$ and phases $\{\varphi_m^{\min}\}$ of the minimum-phase spectrum of the original signal by

$$A_m = \exp\left(c_0 + 2\sum_{n=1}^{N_0} c_n \cos(2\pi nm / L)\right),$$
$$\varphi_m^{\min} = -2\sum_{n=1}^{N_0} c_n \sin(2\pi nm / L), \tag{11}$$

where L is the pitch-period in samples, $1 \le m \le [L/2]$, and $[L/2]$ denotes the integral part of the number $L/2$.

The minimum phases are modified by superimposing the phases of the impulse response of the Hilbert transformer used for excitation in the cepstral synthesizer. Instead of computing spectral flatness and using mixed excitation emphasizing noise at higher frequencies, the maximum voiced frequency f_{max} is used. It is determined from the amplitude spectrum comparing the frequency distances between the pitch frequency multiples and the frequencies of the spectrum local maxima, for voiced segments, and it is zero for unvoiced segments. Then, phases at frequencies higher than f_{max} are randomized. The logarithmic spectrum computed from the truncated cepstrum should represent the speech spectral envelope, but it is vertically shifted towards lower amplitude values. One solution of this problem is the cepstral coefficients determination with gain correction [7] inspired by gain matching in the cepstral speech model [3], however, using different procedures. Peak picking is used to find all the local maxima of the spectrum of the residual signal. It means that all the frequencies at which the spectral slope changes from positive to negative are chosen. Amplitudes at these frequencies are averaged to get the correction gain G. The resulting phases $\{\varphi_m\}$ together with the amplitudes $\{A_m\}$ and the pitch frequency multiples $\{f_m\}$ are used for the final harmonic synthesis.

$$s(l) = 2G\sum_{m=1}^{[L/2]} A_m \cos(2\pi f_m l + \varphi_m), \quad 1 \le l \le L. \tag{12}$$

6. Frequency properties comparison

The comparison of frequency properties for both the approaches has been performed on the stationary parts of the vowels and the consonants (male speaker, $f_0 \approx 110$Hz), with the sampling frequency of 8 and 16 kHz. The smoothed spectra of the resynthesized signals according to both the models have been compared with the smoothed spectrum of the original speech signal. The RMS spectral measure between the spectrum of the synthesized signal and the spectrum of the original signal has been used as a comparison criterion.

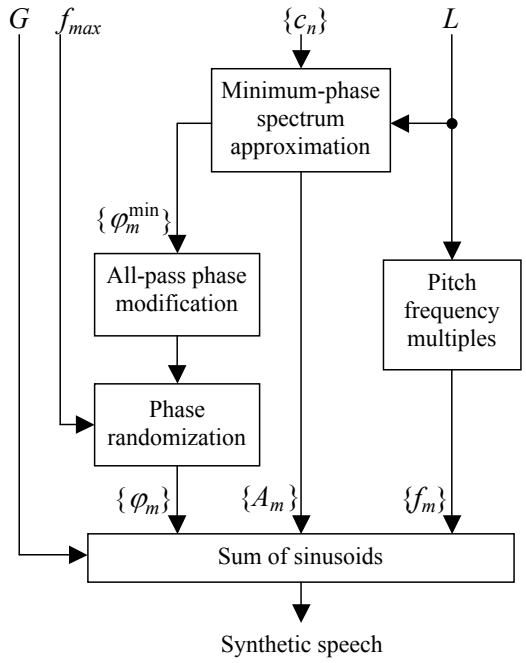

Fig. 5 Block diagram of the harmonic synthesizer with cepstral description.

The RMS values have been computed for the spectra of the speech segments weighted by 24-ms normalized Hamming window, zero-padded to 2048-point FFT for both sampling frequencies. The mean values for about 450 segments have been computed. They contain the error of the excitation (comprised in the phase error of the harmonic model), and the error of the vocal tract model (approximation error) [8]. The spectrum error given by the mean RMS spectral measure of both synthesis methods for 5 vowels, 2 nasals, and 1 fricative is shown in Tab. 2.

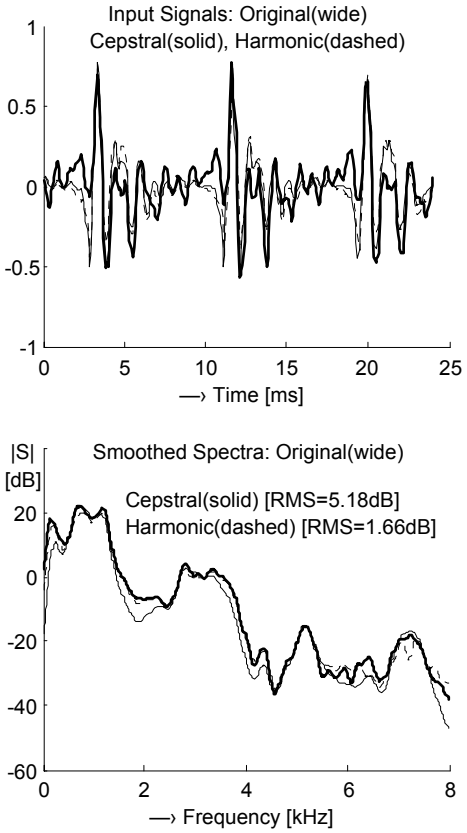

Fig. 6 Comparison in the time and frequency domain – 2nd segment of stationary part of the vowel "a" (f_s=16kHz).

Speech sounds	Mean RMS spectral measure [dB]					
	f_s = 8 kHz			f_s = 16 kHz		
	Ceps. synt.	Harm. synt.	NS*)	Ceps. synt.	Harm. synt.	NS*)
A	3.77	2.37	81	4.29	2.58	78
E	3.69	2.32	60	4.37	2.71	57
I	3.85	2.71	60	4.42	2.60	57
O	4.10	2.71	69	4.48	2.53	69
U	4.62	3.45	60	4.88	3.61	60
M	4.01	2.98	44	4.86	3.14	44
N	4.19	2.97	69	4.81	3.29	66
S	4.68	4.56	10	4.48	4.49	8

*) Number of processed segments

Tab. 2 The mean RMS spectral measure values for the cepstral and harmonic synthesis.

7. Computation complexity

Speech parameterization and synthesis using the cepstral and the harmonic methods were realized in MATLAB program system, which is very suitable for testing and simulation. The computational complexity was compared with the help of the MATLAB function FLOPS (Floating point operation count) and has been referred to one sample of the processed speech signal [9]. The comparison was performed in the following areas:

a) Cepstral speech analysis

- Segment classification (voiced/unvoiced) and pitch-period detection
- Computing coefficients for the cepstral as well as harmonic model

b) Speech synthesis

- Cepstral synthesis
- Computing parameters of the harmonic model from the cepstral coefficients
- Harmonic synthesis

Computational complexity and memory requirements were compared. The memory requirements are important for practical implementation in another programming languages (assembler or C for signal processors). The computational complexity has influence especially on real time applications (speech coders and decoders or text-to-speech systems). The input parameters for the cepstral and har-

monic synthesis at 8 and 16 kHz sampling frequencies are contained in Tab. 3.

Synthesis type / f_s [kHz]	Input parameters necessary for 1 segment synthesis *)	Σ
Cepstral/8	$1\times En$, $25\times\{\tilde{s}_n\}$, $1\times S_F$, $1\times L$	28
Cepstral/16	$1\times En$, $50\times\{\tilde{s}_n\}$, $1\times S_F$, $1\times L$	53
Harmonic/8	$1\times G$, $25\times\{c_n\}$, $1\times f_{max}$, $1\times L$	28
Harmonic/16	$1\times G$, $50\times\{c_n\}$, $1\times f_{max}$, $1\times L$	53

*) Data considered in standard format Integer (length 2 bytes)

Tab. 3 Analysis to synthesis data transfer vector storage requirements.

The mean values of the computational complexity for the whole analysis (including segment classification and pitch-period detection) and synthesis (including computing of parameters of the harmonic model from the cepstral coefficients) operation blocks are summarized in Tab. 4.

Operations	Complexity [FLOPS/sample]	
	8 kHz	16 kHz
Analysis for cepstral synth.	1004.8	1066.3
Analysis for harmonic synth.	1020.8	1069.9
Cepstral synthesis	262.8	404.5
Harmonic synthesis	374	663.7

Tab. 4 Computational complexity.

8. Conclusion

Cepstral description of the vocal tract model transfer function can be used in the source-filter model (denoted as the cepstral speech model in this paper) or in the harmonic speech model with cepstral parameterization. For voiced speech, the source-filter model consists of excitation modeling the glottal activity and filter modeling the vocal tract properties. In the harmonic model the amplitudes and the minimum phases correspond to the magnitude and phase frequency responses of filter modeling the vocal tract. The phase response of the excitation is comprised in all-pass phase modification of the harmonic model. For unvoiced speech, the excitation is represented by random noise corresponding to phase randomization in the harmonic model.

According to the above-mentioned results the harmonic synthesis with cepstral parameterization gives better frequency properties than the cepstral synthesis. Listening tests of the Czech and Slovak isolated words have shown only small audible differences between both the compared synthesis approaches.

The harmonic synthesis based on the cepstral parameterization gives better frequency properties than cepstral synthesis. Results in Tab. 3 and 4 show that storage requirements of both methods are identical, the computational complexity of the analysis is similar, and harmonic synthesis has higher computational complexity than cepstral one.

References

[1] VÍCH, R., SMÉKAL, Z. LPC and CCF Vocal Tract Models in Speech Synthesis. In Proceedings of the 9th European Signal Processing Conference EUSIPCO 98, Rhodes (Greece), 1998, p.1861-1864.

[2] VÍCH, R., PŘIBIL, J., PTÁČEK, M. Cepstrales Sprachsynthesesystem für die tschechische Sprache. In 8. Konferenz Elektronische Sprachsignalverarbeitung, Cottbus (Germany), 1997, p. 218-225.

[3] VÍCH, R. Cepstral Speech Model, Padé Approximation, Excitation and Gain Matching in Cepstral Speech Synthesis. In Proceeding of the 15th Biennial International EURASIP Conference Biosignal'2000, Brno, 2000, p. 77-82.

[4] PŘIBIL, J. Pitch synchronous analysis and synthesis in cepstral speech modelling. In Proceedings of the 10th International Conference RADIOELEKTRONIKA 2000. Bratislava, 2000, p. III 13-16.

[5] MCAULAY, R. J., QUATIERI, T. F. Low-Rate Speech Coding Based on the Sinusoidal Model. Furui, S., Sondhi, M, M, Eds.: Advances in Speech Signal Processing, Marcel Dekker, New York, 1992.

[6] OPPENHEIM, A., V., SCHAFER, R., W. Discrete-Time Signal Processing. Prentice Hall, 1989.

[7] MADLOVÁ, A. Harmonic Speech Model with Cepstral Parameterization. VÍCH, R., ed. SPEECH PROCESSING, 10th Czech-German Workshop, Prague, 2000, p. 56-58.

[8] MADLOVÁ, A., PŘIBIL, J. Comparison of two approaches to speech modelling based on cepstral description. In Proc. of the 15th Biennial International EURASIP Conference Biosignal 2000, Brno, 2000, p. 83-85.

[9] PŘIBIL, J., MADLOVÁ, A. Computational complexity of two methods based on cepstral parameterization of speech signal. In Proc. of the 5th International Conference New Trends in Signal Processing, Liptovský Mikuláš (Slovakia), 2000, p. 248-251.

About authors...

Jiří PŘIBIL was born in 1962 in Prague, Czechoslovakia. He received Ing (MSc) degree in computer engineering and PhD degree in applied electronics from the Faculty of Electrical Engineering, Czech Technical University Prague in 1991 and 1998, respectively. Now he is scientist in Dept. of Digital Signal Processing, in Inst. of Radio Engineering and Electronics of the Czech Academy of Sciences. His research interests are signal processor and applications, speech analysis and synthesis, text-to-speech systems.

Anna MADLOVÁ was born in Hlohovec, Czechoslovakia in 1962. She received Ing (MSc) degree in radio electronics (medical electronics) from the Slovak Technical University in Bratislava in 1985. For six years she had been with Chirana Research Center for Medical Equipment as a research assistant. Since 1992 she has been working as a university teacher at the Dept. of Radio Electronics, Slovak University of Technology in Bratislava.